Distributed Information Processing

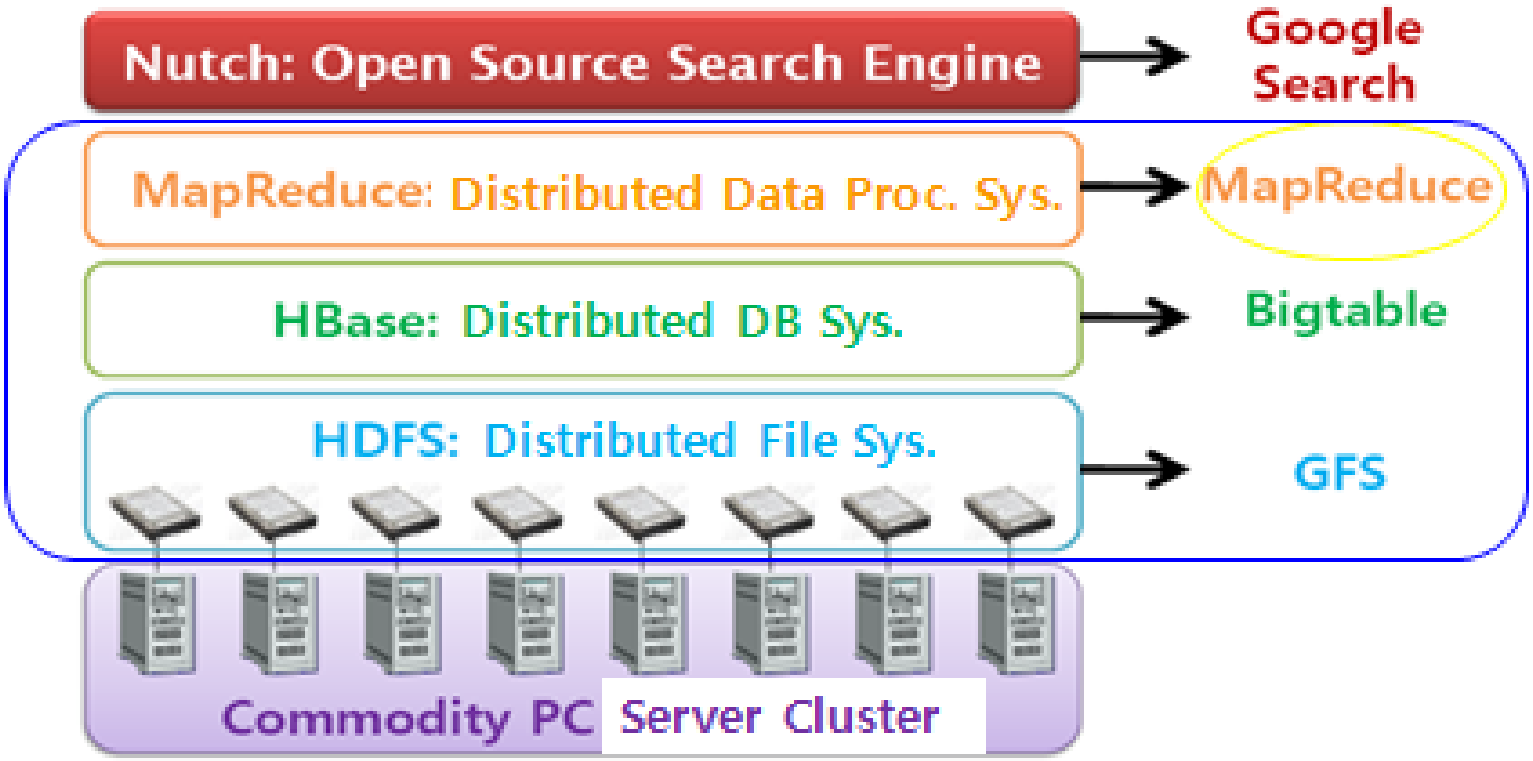
22nd Lecture

Eom, Hyeonsang (엄현상)
Department of Computer Science & Engineering
Seoul National University

Outline

- Hadoop & MapReduce Example
- Q&A

Hadoop Architecture



MapReduce Overview

- MapReduce: a Framework for processing huge datasets stored in a filesystem or database using a large number of computers (nodes)
  - Map step: The master node takes the input, partitions it up into smaller sub-problems, and distributes those to worker nodes. The worker node processes that smaller problem, and passes the answer back to its master node
  - Reduce step: The master node then takes the answers to all the sub-problems and combines them in some way to get the output

Wikipedia

MapReduce Overview Cont'd

- Map: Map(k1, v1)->list(k2, v2)
  - All independent maps performed in parallel
    - Limitation: data source and the number of CPUs
- Reduce: Reduce(k2, list(v2))->list(v3)
  - Each reducer presented outputs of the map operation sharing the same key
- Parallelism leading to fault tolerance
- Connecting the processes
  - Distributed file system
  - Direct streaming

Wikipedia

MapReduce Overview Cont'd

- Data flow
  - Input reader
    - Input from stable storage -> splits (possibly of key/value pairs)
  - Map function
    - Splits -> key/value pairs
  - Partition function
    - Key and # of reducers -> index of the desired reduce

Wikipedia

Map Reduce Overview Cont'd

- Data flow cont'd
  - Shuffle
    - Parallel sort & exchange
      - Transient data usually stored on local disk and fetched remotely by the reducers
  - Comparison function
    - Sorting the input for each reduce using the func.
  - Reduce function
    - Sorted keys with values -> those with reduced values
  - Output writer
    - Output to stable storage

Wikipedia

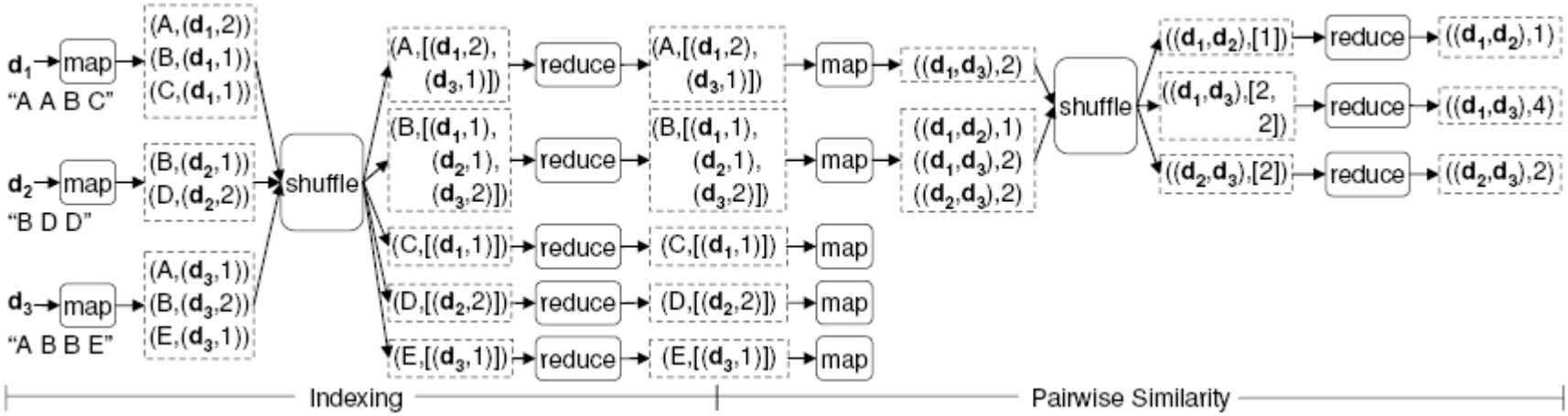
Document Clustering

- Pairwise Document Similarity

$$sim(d_i, d_j) = \sum_{t \in V} w_{t,d_i} \cdot w_{t,d_j}$$

where $sim(d_i, d_j)$ is the similarity between the documents, and $w_{t,d}$ indicates the importance of each term t in the documents

MapReduce Example: Document Clustering



Pairwise Document Similarity in Large Collections with MapReduce